Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15655
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Forecasting Hotel Demand with Time Series Prediction Model Using Random Forest Regression

Dewa Ayu Kadek Pramita[1)], Ni Wayan Sumartini Saraswati[2)*], I Putu Dedy Sandana[3)], Dewa Ayu Putu Rasmika Dewi[4)], Ni Kadek Bumi Krismentari[5)]
[1,2,3,5)]Institut Bisnis dan Teknologi Indonesia, Indonesia, [4)]Monash University, Australia
[1)]pramita.wayu@instiki.ac.id, [2)]sumartini.saraswati@gmail.com, [3)]dedy.sandana@instiki.ac.id, [4)]mika.dewi@monash.edu, [5)]kadek_bumi@instiki.ac.id

Abstract: The tourism sector, as one of the main contributors to national foreign exchange, relies heavily on the growth of the hospitality industry. Improvements in this sector are expected to enhance service quality and strengthen the overall image of tourism. However, the hospitality industry is highly dynamic, with fluctuating room demand influenced by both internal and external factors, creating challenges for accurate demand forecasting. This study develops a hotel demand prediction model using internal variables (occupancy rate, reservations, cancellations, and lead time) and external variables (events and visitor numbers). The Random Forest Regression method was employed, with predictive performance evaluated through a proxy demand index. The dataset was obtained from Adiwana Unagi Suites, Ubud, Bali, covering historical time series data from November 2021 to July 2025 with a total of 18.674 transactions. Evaluation metrics included Mean Absolute Error, Mean Square Error, Root Mean Square Error, and R-squared, applied to each hotel room type. The results demonstrate strong predictive performance, with R-squared values of 99.83% for test data, 99.95% for training data, and 88.24% for three-month prediction data, accompanied by low error values across all metrics. The lower performance in the three-month forecast may be due to the proxy demand index not fully representing actual demand. Overall, the findings highlight the potential of machine learning approaches, particularly Random Forest Regression, to support decision-making in hotel management. The model can serve as a reference for room pricing, allocation, and operational strategies, enabling stakeholders to adapt effectively to fluctuating market demand.

Keywords: Hotel Demand; Prediction; Random Forest Regression; Reservation; Time Series.

INTRODUCTION

The hotel industry is a highly dynamic and competitive sector (George & Mishra, 2024), where hotel management performance strongly depends on room demand (Sampaio et al., 2024). Hotel room demand fluctuates according to various factors such as seasonality, travel trends, and location. To improve operational efficiency and business strategies, hotels require accurate demand forecasting models (Shirisha et al., 2023). Such models can assist hotels in setting optimal room prices, avoiding prices that are too low (leading to revenue loss) or too high (potentially driving away customers). Forecasting demand indices also helps management in planning marketing and distribution strategies, preventing both overbooking and excessive vacancies while maintaining operational efficiency and staff management (Gomez-Talal et al., 2025). Most existing hotel demand prediction models still rely heavily on internal historical data, such as past occupancy and reservations. However, external factors such as events, holiday seasons, and visitor arrivals also influence hotel demand. Large-scale events can immediately trigger surges in hotel demand, while national holidays often result in significant increases in bookings. Similarly, visitor arrival numbers provide an overall scale of demand, where higher arrivals indicate higher potential hotel demand. This study develops a daily hotel demand prediction model using both internal and external variables within a machine learning framework. Internal variables are derived from hotel reservation processes, including occupancy rate, number of bookings, cancellations, and lead time. External data can be sourced online, such as visitor arrival statistics and event calendars. In this study, model evaluation is based on a proxy variable approach. Several key internal variables, such as number of bookings, occupancy rate, and

*name of corresponding author

cancellations, are selected as demand representations, with weights assigned according to their influence. Thus, the resulting demand index reflects hotel demand more comprehensively than relying on a single indicator.

Several studies on tourism demand forecasting have been conducted. For instance, (Abdou et al., 2022) applied Random Forest and Linear Regression with monthly data, showing that Random Forest outperformed Linear Regression, achieving R-squared of 72% on test data and 95% on training data. Study (Hewapathirana, 2025) compared machine learning models (Support Vector Regression, Random Forest, and Artificial Neural Networks) with the traditional time-series Seasonal Autoregressive Integrated Moving Average (SARIMA) for tourist demand forecasting in Sri Lanka, using monthly tourist arrival data. It also incorporated external data from TripAdvisor forums and Google Trends. Based on MAE, MSE, and MAPE evaluations, machine learning consistently outperformed traditional time-series models. Study (Laaroussi et al., 2023) proposed a hybrid deep learning framework combining search query data, autoencoders, and stacked LSTM to forecast monthly tourist arrivals in Marrakech, Morocco. The proposed model achieved an R-squared of 97.45%, with the best performance obtained from the AE-3 Layer LSTM. A study in South Korea (Kim et al., 2021) predicted daily tourism demand using a multi-head attention CNN (MHAC) model, integrating external factors such as politics, diseases, seasons, and attractions (Google Trends). The model achieved a MAPE of 25.7%, outperforming 1D-CNN, Bi-LSTM, and CNN-LSTM. In the tourism transportation domain, (Peng et al., 2021) employed BERT and Gradient Boosting Regression Trees (GBRT) to forecast tourist passenger demand using variables such as social network posts (Weibo), weather, holidays, and historical passenger data. Text was processed with BERT to extract sentiment scores, followed by forecasting with GBRT, achieving a MAPE of 4.74%, outperforming baseline methods.

Other studies have specifically focused on daily hotel demand. Study (B. Zhang et al., 2022) for hotels in Sanya, Hainan, used daily search frequency data to improve monthly hotel demand forecasting. A hybrid mixed-data sampling (MIDAS) regression integrating dynamic factors showed improved accuracy, with the best model (MIDAS-C) achieving an R value of 0.9698 on the Sanya dataset. Study (Ampountolas, 2021) implemented time-series models to predict daily hotel demand across multiple horizons in US hotels, incorporating exogenous variables such as temperature, holidays, hotel competitiveness, and weekdays. Results showed that sGARCH and GJR-GARCH outperformed alternative models across all horizons. Study (H. Zhang & Lu, 2022) integrated econometric modeling with scenario analysis to capture uncertainty caused by COVID-19. Using ARDL-ECM enhanced with compound scenario analysis, including travel restrictions, vaccination rates, government policies, and healthcare resilience, the study demonstrated improved accuracy compared to conventional econometric models. Study (Huang & Zheng, 2021) proposed a deep learning model with spatial and temporal correlations, incorporating agglomeration effects, attention mechanisms, and Bayesian optimization. Using historical daily hotel demand data in Xiamen, China, the model significantly outperformed benchmarks, achieving RMSE of 46.008. Another study in Vienna (Gunter, 2021) developed a hotel demand model based on seasonality, long-term trends, and internal variables such as ADR. Models applied included Seasonal Naïve, ETS, SARIMA, TBATS, and Seasonal NNAR, with evaluation conducted via rolling windows for 1 - 90 day forecasts. Bates-Granger weights and ranks produced the best overall performance.

From prior research, most studies have focused on tourism demand rather than hotel-specific demand. Many also rely on monthly data, with fewer using daily data and often limited to internal historical reservations. Integration of external factors such as event calendars and visitor numbers remains scarce. Methodologically, several studies suggest Random Forest Regression outperforms other approaches. However, its application to daily hotel demand prediction combining internal and external variables has not yet been fully explored. Therefore, this study addresses this gap by developing a hotel demand prediction model using Random Forest Regression that integrates internal variables (occupancy, reservations, cancellations, and lead time) and external variables (event calendars and visitor arrivals). Model performance is evaluated using RMSE, MAE, MSE, and R-squared. This approach is expected to produce more accurate predictions. Practically, the research contributes to providing a forecasting system that supports daily operational decision-making in hotels. Academically, it fills a gap in tourism forecasting literature by offering a multi-source data approach

LITERATURE REVIEW

Research on hotel and tourism demand forecasting has expanded significantly as the hospitality industry increasingly relies on predictive analytics to optimize revenue management strategies. Recent studies demonstrate that machine learning approaches can outperform traditional statistical models when dealing with complex, non-linear, and dynamic patterns in booking behavior. (Shirisha et al., 2023) show that machine learning algorithms can effectively predict hotel bookings and cancellations by learning from historical data patterns, highlighting the capability of ML models to capture interactions that linear models may overlook.

Similarly, (Gomez Talal et al., 2025) develop interpretable machine learning models for forecasting hotel booking cancellations, emphasizing not only predictive accuracy but also model transparency. Their work underscores the importance of techniques that allow decision-makers to identify key features influencing forecasts.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15655
e-ISSN : 2541-2019
p-ISSN : 2541-044X

This aligns closely with the strengths of Random Forest Regression, which provides interpretability through feature importance analysis.

In a broader tourism context, (Hewapathirana, 2025) evaluates various machine learning models for tourism demand forecasting, incorporating social media data to enhance predictive performance. The study demonstrates that machine learning models are highly adaptive to evolving traveler behavior, especially when leveraging multi-source datasets. This supports the use of Random Forest, which excels in handling diverse nonlinear features in forecasting tasks.

Deep learning approaches have also shown promising results. (Laaroussi et al., 2023) propose a hybrid deep learning framework that significantly improves forecast accuracy. However, such models require larger datasets and higher computational resources, making Random Forest a more practical alternative in scenarios with limited data or operational constraints.

(Peng et al., 2021) further highlight the value of integrating digital and social network variables into tourism demand forecasting, showing strong correlations between online search behaviors and actual tourism flows. These findings suggest that hotel demand forecasting could benefit from incorporating similar external features.

(Ampountolas, 2021) compares SARIMAX, neural networks, and GARCH models for daily hotel demand forecasting, demonstrating that model performance depends heavily on data characteristics and forecast horizons. This provides an important basis for evaluating how machine learning models such as Random Forest compare to traditional time series approaches.

A synthesis of these studies reveals several research gaps. First, despite extensive work on deep learning and hybrid models, few studies have specifically explored the performance of Random Forest Regression in the context of hotel demand forecasting. Second, much of the current literature focuses on predicting booking cancellations rather than forecasting overall hotel demand. Third, interpretability remains underexplored in time series ML forecasting for hospitality. These gaps provide clear motivation for the present research, which aims to evaluate Random Forest Regression as an accurate and interpretable method for forecasting hotel demand.

METHOD

The research flowchart is presented in Figure 1. This study utilizes historical data from Adiwana Unagi Suites, a hotel located in Ubud, Gianyar, Bali. The historical dataset was obtained from the hotel's Property Management System (PMS), covering the period from November 2021 to July 2025. The hotel comprises five room types: Adiwana Forest View, Adiwana Pool Access, Club Room, One-Bedroom Pool Villa, and Unagi Suites. A separate predictive model is developed for each of these room types. For external data, this study employs tourist arrival statistics obtained from the official websites of the Central Bureau of Statistics (Badan Pusat Statistik, BPS) and the Bali Provincial Tourism Office. National holiday data were retrieved from the website of the Coordinating Ministry for Human Development and Cultural Affairs, while major events in Bali, particularly in Ubud, were collected from the Bali Provincial Tourism Office website. In total, the historical hotel dataset consists of 18,674 reservation records. A sample of the historical dataset is presented in Table 1.

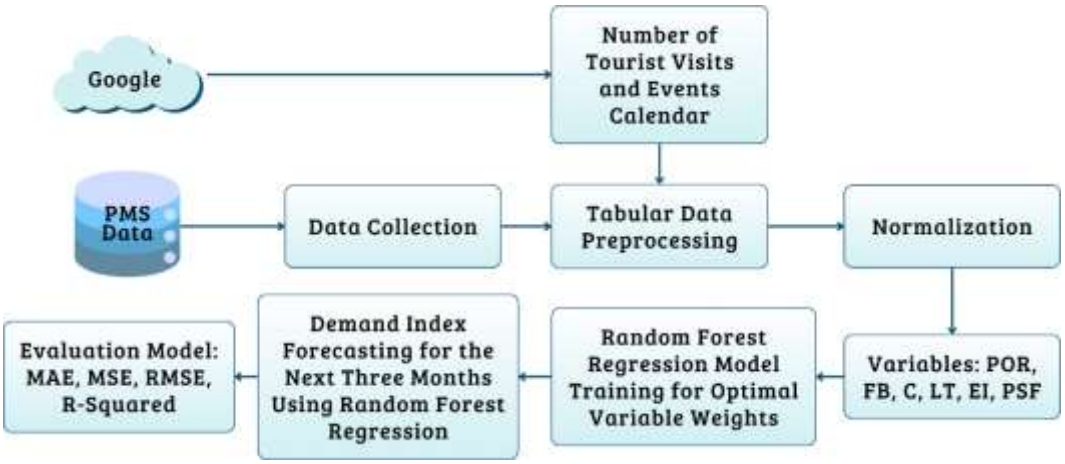


Figure 1. Research Workflow Diagram

Table 1. Sample of Historical Internal Data

Date	Room Type	Lead Time	Check in	Check out	Status
2021-11-01	Adiwana Forest View	7	2021-11-08	2021-11-10	Cancelled
2021-11-01	Adiwana Forest View	20	2021-11-21	2021-11-22	Cancelled

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15655
e-ISSN : 2541-2019
p-ISSN : 2541-044X

2021-11-02	Adiwana Forest View	3	2021-11-05	2021-11-06	Cancelled
2025-07-26	Unagi Suites	18	2025-08-13	2025-08-16	Definite
2025-07-26	Unagi Suites	44	2025-09-08	2025-09-10	Definite

The data undergo a cleaning process to address missing values and duplicates. Missing values are handled through imputation, while duplicate records are removed to retain only unique entries. Data preprocessing aims to generate a daily data summary, which is achieved by aggregating transaction records that fall on the same date based on daily intervals and room type. Similarly, visitor arrival data, originally in monthly format, are transformed into daily data. After preprocessing, the total number of daily records is 1,366 rows. Subsequently, the tabular dataset is normalized into percentage form. This study employs hotel occupancy, number of bookings, and number of cancellations as a proxy demand index (PDI) to substitute for the actual demand index, with each variable weighted according to its relative influence on demand. The index is formulated as a weighted linear combination of these variables, expressed as shown in Equation (1):

$$Proxy\ Demand\ Index = w_1 \times Bookings + w_2 \times Occupancy - w_3 \times Cancellations \quad (1)$$

where w_1, w_2, and w_3 denote the relative weights of each component. The weights are not assigned heuristically but are derived from the feature importance values produced by the Random Forest Regression model, reflecting the empirical contribution of each variable to demand prediction. From this preprocessed dataset, the variables used to develop the hotel Demand Index (DI) prediction model are constructed, namely:

a. Predicted Occupancy Rate (POR)
   This variable is developed by predicting future hotel occupancy rates based on historical occupancy data. The method employed is time-series prediction using polynomial regression, as applied in a previous study conducted in 2024 (Pramita et al., 2024).
b. Forward Booking (FB)
   This variable represents the number of reservations made for future dates in comparison to the maximum reservation capacity.
c. Cancellation (C)
   This variable represents the number of reservations canceled by customers within a specific period in comparison to the total number of reservations made.
d. Lead Time (LT)
   This variable represents the number of days between the booking date and the guest's check-in date, divided by the maximum lead time of three months.
e. Event Impact Factor (EI)
   This variable represents the impact scale of major events or national holidays.
f. Predicted Seasonality Factor (PSF)
   This variable is developed from the recognition of tourist arrival patterns to Bali using a regression method based on visitor arrival data from BPS.

2.1 Random Forest Regression

Random Forest Regression is employed to predict the hotel demand index. Random Forest combines multiple decision trees with the objective of improving prediction accuracy while reducing overfitting (Thomas & Kaliraj, 2024). Each decision tree is trained independently on different subsets of the data, and the results are then averaged or aggregated through voting to generate the final prediction (Han et al., 2021). Random Forest can be applied to both classification and regression tasks. While both share the same principle of bagging, they differ in terms of outputs and aggregation methods. In regression, Random Forest produces continuous values, and the aggregation function is based on averaging (K et al., 2024).

In this study, a negative correlation (inverse relationship) is implemented for variable C. Negative correlation indicates features that move in opposite directions. Within the context of this research (variable C), it is assumed that the higher the cancellation rate, the lower the effective demand for hotel rooms. This assumption aligns with the findings of (Hikmawati et al., 2024), which highlight that a high cancellation rate negatively impacts room utilization and hotel revenue. Consequently, cancellations not only affect capacity utilization and revenue performance but also indirectly reduce actual demand.

Therefore, variable C is transformed into negative values before being fed into the model. Considering that demand for each room type may influence its pricing, separate models are developed for each room type.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15655
e-ISSN : 2541-2019
p-ISSN : 2541-044X

2.2 Model Evaluation

The regression model in this study is evaluated using MAE, MSE, RMSE, and R-squared (R^2). Model evaluation used a time-based (chronological) data split rather than random sampling. The first 80 percent of observations were used for training, while the remaining 20 percent were used for testing, ensuring that future information did not leak into the training process. In addition, model performance was assessed using a three-month out-of-sample forecasting period to evaluate generalization to future demand. Out of a total of 1,366 daily records, the training dataset consists of 1,092 rows, while the testing dataset includes the last 274 rows. R-squared is employed to measure how well the model explains data variability (Hossain et al., 2025). The value ranges from 0 to 1, where 0 indicates that the model is unable to explain data variance, while 1 indicates that the model fully explains all variance in the data. In general, the higher the R-squared value, the better the model fits the data (Soegianto et al., 2024). Mathematically, R-squared can be expressed as shown in Equation (2).

$$R^2 = 1 - \frac{RSS}{TSS} \tag{2}$$

Residual Sum of Squares (RSS) represents the total squared difference between the actual and predicted values (Karawapong et al., 2025). In other words, it measures how far the predictions deviate from the actual data. Total Sum of Squares (TSS) indicates the overall variation in the data relative to the mean, representing the total variability in the dataset. RSS and TSS are defined in Equation (3) and Equation (4), where $\bar{y}$ denotes the mean of the actual data points. Mean Squared Error (MSE) is the average of the squared differences between the actual and predicted values. MSE is always non-negative (Ray et al., 2023). Root Mean Squared Error (RMSE) is the square root of MSE. Unlike MSE, RMSE is expressed in the same unit as the original target variable, making it more interpretable (Rahmaddeni et al., 2024). Mean Absolute Error (MAE) calculates the average of the absolute differences between the actual and predicted values (Natarajan et al., 2024). MSE, RMSE, and MAE are represented in Equations (5) – (7), where $\hat{y}$ denotes the predicted value.

$$RSS = \sum_{i=1}^{n} (y_i^{Actual} - y_i^{Predicted})^2 \tag{3}$$

$$TSS = \sum (y_i - \bar{y}_i)^2 \tag{4}$$

$$MSE = \frac{RSS}{n} = \sum_{i=1}^{n} (y_i^{Actual} - y_i^{Predicted})^2 / n \tag{5}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{RSS}{n}} = \sqrt{\sum_{i=1}^{n} (y_i^{Actual} - y_i^{Predicted})^2 / n} \tag{6}$$

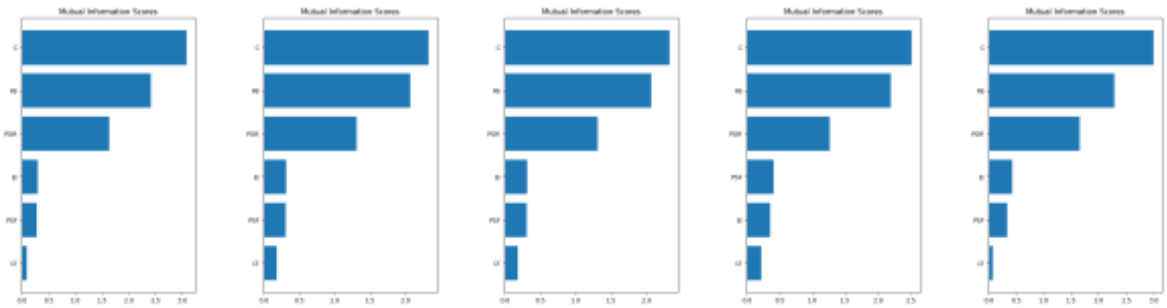
$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{7}$$

RESULT

To enhance model interpretability, mutual information scores were used to assess the contribution of each explanatory variable to hotel room demand prediction (Figure 2). The analysis was conducted separately for each room type to capture differences in demand drivers. The results indicate that cancellations (C) exhibit the highest mutual information scores, followed by POR and FB, suggesting that short-term booking volatility plays a dominant role in shaping demand predictions.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15655
e-ISSN : 2541-2019
p-ISSN : 2541-044X



(a) (b) (c) (d) (e)

Figure 2. Mutual Information Score (a) Adiwana Forest View, (b)Adiwana Pool Access, (c)Club Room, (d)One Bedroom Pool Villa, (e) Unagi Suites

The sample dataset for one of the room types (Adiwana Forest View), after undergoing cleaning, preprocessing, and normalization to generate the variables required for the model, is presented in Table 2. The evaluation results after building the Random Forest Regression model are shown in Table 3. Based on the evaluation results from the test, train, and prediction datasets, the highest R-squared was achieved at 99.95% on the training data of Model 5 (Unagi Suites), while the lowest R-squared was 73.53% on the three-month prediction data of Model 1 (Adiwana Forest View). The models demonstrate strong performance on both training and testing datasets, as all R-squared values exceeded 80%, with test data ranging from 88.76% to 99.83% and training data ranging from 84.47% to 99.95%. These results indicate that the models on training and testing datasets are highly capable of explaining historical data variation. For the three-month prediction dataset, however, two models yielded R-squared values below 80%, with the overall range falling between 73.53% and 88.24%.

Table 2. Sample of Model Variables for Adiwana Forest View Room Type

Date	POR	C	EI	PSF	LT	FB
2025-07-24	0.875	-0.262500	1.017857	0.44779	0.033067	1.250
2025-07-25	1.000	-0.111111	1.017857	0.44779	0.116279	1.125
2025-07-26	1.125	-0.204545	1.017857	0.44779	0.130814	1.375
2025-07-27	1.000	-0.200000	1.017857	0.44779	0.116764	1.250
2025-07-28	1.125	-0.112500	1.125000	0.44779	0.088299	1.250

Table 3. Sample of Model Variables for Adiwana Forest View Room Type

Data	Model Random Forest Regression	Room Type	MAE	MSE	RMSE	R-Squared
Data Test	1	Adiwana Forest View	0,0379	0,0044	0,0664	91,35%
	2	Adiwana Pool Access	0,0033	0,0002	0,0123	99,63%
	3	Club Room	0,0088	0,0004	0,0204	99,34%
	4	One Bedroom Pool Villa	0,0639	0,0074	0,0858	88,76%
	5	Unagi Suites	0,0016	0,00004	0,0078	99,83%
Data Train	1	Adiwana Forest View	0,0502	0,0104	0,1018	84,47%
	2	Adiwana Pool Access	0,0023	0,0001	0,0117	99,81%
	3	Club Room	0,0055	0,0003	0,0180	99,79%
	4	One Bedroom Pool Villa	0,0698	0,0117	0,1081	87,94%
	5	Unagi Suites	0,0015	0,00006	0,0064	99,95%
Data Prediksi (3 bulan ke depan)	1	Adiwana Forest View	0,0895	0,0170	0,1304	73,53%
	2	Adiwana Pool Access	0,0469	0,0038	0,0619	88,24%
	3	Club Room	0,0818	0,0146	0,1206	80,41%
	4	One Bedroom Pool Villa	0,0896	0,0136	0,1166	77,59%
	5	Unagi Suites	0,0470	0,0052	0,0721	86,76%

Between the training and testing datasets, a slight decrease in performance was observed for Models 2, 3, and 5. Conversely, in Models 1 and 4, the testing results were slightly higher than the training results, though the gap was not substantial. This phenomenon may occur because the training dataset is much larger than the testing dataset, making the training data more complex and diverse. In comparison with the three-month prediction data, a more significant decrease in performance or a larger gap was evident. Based on historical data training, the Random Forest Regression models incorporated seasonal factors and historical data patterns. However, the Proxy Demand Index (PDI) used as the evaluation metric does not yet fully reflect the actual hotel room demand for the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15655
e-ISSN : 2541-2019
p-ISSN : 2541-044X

three-month prediction period, as the PDI values are still provisional. Since not all tourists have made reservations in advance, the predictive performance for the three-month horizon tends to be lower than that of historical data. Models 2, 3, and 5 still performed reasonably well in the three-month prediction, as their R-squared values remained above 80%. However, the significant drop in performance for Models 1 and 4 may indicate overfitting, where the models are not sufficiently robust for forward-looking predictions. The evaluations of RMSE, MSE, and MAE were consistent with R-squared results across all datasets, where higher R-squared values were always associated with lower RMSE, MSE, and MAE values. Since the evaluation metrics for all models approached zero, this suggests that the models were capable of capturing the variability within the data.

When visualized using line charts, the PDI (Actual) and the three-month prediction data are shown in Figures 3 -7. Across Models 1 to 5, the prediction lines follow the general pattern of the actual PDI line. However, in Model 1 (Adiwana Forest View), a stagnation occurred from mid-September to late October, where the model failed to capture the pattern within that range. This limitation is reflected in its lower R-squared value. Based on the evaluations and line chart visualizations, the room types Adiwana Pool Access, Club Room, and Unagi Suites were relatively stable in tracking the PDI trend, demonstrating greater accuracy and robustness. Meanwhile, Adiwana Forest View and One-Bedroom Pool Villa appeared more volatile and require closer monitoring. For Adiwana Forest View in particular, predictions may be more reliable when applied with a shorter forecasting horizon.

Figure 3. Actual and Prediction Room Type Adiwana Forest View

Figure 4. Actual and Prediction Room Type Adiwana Pool Access

Figure 5. Actual and Prediction Room Type Club Room

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15655
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Figure 6. Actual and Prediction Room Type One Bedroom Pool Villa

Figure 7. Actual and Prediction Room Type Unagi Suites

The findings of this study, when compared to previous research, demonstrate several differences. In terms of methodology, the Random Forest Regression applied in this study achieved higher R-squared values than those reported in (Abdou et al., 2022), indicating that Random Forest Regression was able to capture demand variation more effectively. The combination of internal and external variables in this study produced relatively low RMSE, MAE, and MSE values, alongside R-squared values ranging from 73% to 99.95%. This aligns with studies that incorporated external variables, such as (Peng et al., 2021) (B. Zhang et al., 2022) (H. Zhang & Lu, 2022). Although the model in this research remains relatively simple compared to more complex approaches such as hybrid or deep learning methods used by (Kim et al., 2021) and (Peng et al., 2021) , the results show that even a simpler approach can still deliver strong predictive performance.

These findings provide practical implications for hotel stakeholders. A daily demand prediction model using Random Forest Regression can support strategic decision-making, including room pricing strategies, room allocation, and staff management adjustments during periods of high or low demand. Room types identified as more sensitive to demand fluctuations may require special attention in order to optimize occupancy and hotel revenue. From an academic perspective, this study highlights that Random Forest Regression can yield strong performance for daily hotel demand forecasting. Moreover, the careful selection of relevant external variables is shown to enhance model accuracy. These insights may serve as a foundation for future research, such as comparing machine learning and deep learning methods for more complex demand forecasting, or incorporating regularization techniques to further improve performance. Despite its promising results, this study also has limitations. The external variables used were limited to events and tourist arrival numbers, without including other potentially influential factors such as weather conditions or additional exogenous variables.

DISCUSSIONS

The Random Forest Regression models demonstrated strong performance across most room types, as indicated by high R-squared values and low MAE, MSE, and RMSE on both training and testing datasets. Room types such as Adiwana Pool Access, Club Room, and Unagi Suites exhibited stable demand patterns, allowing the model to generate highly accurate predictions. In contrast, Adiwana Forest View and One-Bedroom Pool Villa showed greater error variation, suggesting that more volatile demand patterns are harder for the model to capture consistently.

A noticeable decline in accuracy occurred in the three-month forecast, primarily due to the Proxy Demand Index (PDI), which does not fully represent actual future demand since forward bookings are not yet complete. This limits the model's ability to replicate true future behavior. Overall, the findings indicate that Random Forest performs very well for room types with stable historical patterns but is less optimal for categories with higher variability. Therefore, forecasting methods and supporting variables should be aligned with the unique characteristics of each room type to maximize the accuracy of the Hotel Demand Index.

*name of corresponding author

CONCLUSION

The results of this study demonstrate that the Random Forest Regression model delivers strong performance across different hotel room types, with the highest R-squared values achieved at 99.83% for test data, 99.95% for training data, and 88.24% for the three-month prediction data. The model successfully captured demand variation, as reflected in the low RMSE, MSE, and MAE values, as well as in predictions that followed the PDI trend in the visualizations. Room types with stable demand, such as Adiwana Pool Access, Club Room, and Unagi Suites, exhibited more robust predictions, whereas Adiwana Forest View and One-Bedroom Pool Villa were more volatile, showing lower predictive performance and stagnation in certain periods. The decline in performance for the three-month prediction data was influenced by the PDI not fully representing actual demand. These findings have important implications for stakeholders, providing a basis for strategic decision-making, such as room pricing, room allocation, and operational management. Although the Random Forest Regression method applied in this study is relatively simple, it produced competitive results compared to studies employing hybrid or deep learning methods. Despite achieving high performance, this study is limited to room types in a single hotel. Future research could expand by comparing alternative methods (hybrid, machine learning, deep learning, and regularization techniques), incorporating additional external variables, and exploring shorter forecasting horizons to further enhance predictive accuracy.

ACKNOWLEDGMENT

We thank Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi for the support, both morally and financially, for the successful implementation of this research. We also express our deepest gratitude to the SinkrOn journal for agreeing to publish the results of our research. We also thank Adiwana Unagi Suites for providing the data needed for this research. Hopefully, this research can be useful for future research

REFERENCES

Abdou, M., Musabanganji, E., & Musahara, H. (2022). Determinants of Tourism Demand Using Machine Learning Techniques. *African Journal of Hospitality, Tourism and Leisure*, *11*(2), 770–780. https://doi.org/10.46222/ajhtl.19770720.256

Ampountolas, A. (2021). Modeling and Forecasting Daily Hotel Demand: A Comparison Based on SARIMAX, Neural Networks, and GARCH Models. *Forecasting*, *3*(3), 580–595. https://doi.org/10.3390/forecast3030037

George, R., & Mishra, V. P. (2024). Analysis and Impact of Digital Influence in Hospitality and Tourism Industry. *2024 2nd International Conference on Sustaining Heritage: Embracing Technological Advancements (ICSH)*, 41–45. https://doi.org/10.1109/ICSH62408.2024.10779719

Gomez Talal, I., Ballesteros, P., & Singh, A. (2025). Machine Learning in Hospitality: Interpretable Forecasting of Booking Cancellations. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2025.3536094

Gunter, U. (2021). Improving Hotel Room Demand Forecasts for Vienna across Hotel Classes and Forecast Horizons: Single Models and Combination Techniques Based on Encompassing Tests. *Forecasting*, *3*(4), 884–919. https://doi.org/10.3390/forecast3040054

Han, S., Williamson, B. D., & Fong, Y. (2021). Improving random forest predictions in small datasets from two-phase sampling designs. *BMC Medical Informatics and Decision Making*, *21*(1), 1–9. https://doi.org/10.1186/s12911-021-01688-3

Hewapathirana, I. U. (2025). Advancing tourism demand forecasting in Sri Lanka: evaluating the performance of machine learning models and the impact of social media data integration. *Journal of Tourism Futures*, *11*(2), 261–285. https://doi.org/10.1108/JTF-06-2023-0149

Hikmawati, N. K., Ramdhani, Y., & Wartika. (2024). Exploring ADR Trends: A Data Mining Approach to Hotel Room Pricing, Cancellations, and EDA. *Journal of Applied Data Sciences*, *5*(1), 189–202. https://doi.org/10.47738/jads.v5i1.165

Hossain, M. F., Das, D., Sultana, F., Istiaque, A., & Hossain, M. A. (2025). Performance Analysis of Machine Learning Models for Predicting Return Loss in 5G Microstrip Patch Array Antenna Design. *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–6. https://doi.org/10.1109/ECCE64574.2025.11013986

Huang, L., & Zheng, W. (2021). Novel deep learning approach for forecasting daily hotel demand with agglomeration effect. *International Journal of Hospitality Management*, *98*(April), 1–11. https://doi.org/10.1016/j.ijhm.2021.103038

K, M. M. M., B, I., Prasad, H., & TD, S. (2024). Load Forecasting Using Random Forest Regression Algorithm in Machine Learning. *2024 International Conference on Science Technology Engineering and Management (ICSTEM)*, 1–6. https://doi.org/10.1109/ICSTEM61137.2024.10560982

Karawapong, A., Karoonsoontawong, A., & Kanitpong, K. (2025). Exploring the multiscale relationship between the built environment and metro station ridership. *Case Studies on Transport Policy*, *20*(April), 101466.

*name of corresponding author

https://doi.org/10.1016/j.cstp.2025.101466

Kim, D. K., Shyn, S. K., Kim, D., Jang, S., & Kim, K. (2021). A Daily Tourism Demand Prediction Framework Based on Multi-head Attention CNN: The Case of the Foreign Entrant in South Korea. *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings*. https://doi.org/10.1109/SSCI50451.2021.9659950

Laaroussi, H., Guerouate, F., & Sbihi, M. (2023). A novel hybrid deep learning approach for tourism demand forecasting. *International Journal of Electrical and Computer Engineering*, *13*(2), 1989–1996. https://doi.org/10.11591/ijece.v13i2.pp1989-1996

Natarajan, E., Radvar, T., Solihin, M. I., Ang, C. K., & Kumar, K. (2024). Chapter 15 - A pilot study and development of prediction model for tire compound quality. In K. Palanikumar, E. Natarajan, S. Ramesh, & J. P. B. T.-M. I. in M. E. Davim (Eds.), *Woodhead Publishing Reviews: Mechanical Engineering Series* (pp. 299–311). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-443-18644-8.00020-4

Peng, T., Chen, J., Wang, C., & Cao, Y. (2021). A Forecast Model of Tourism Demand Driven by Social Network Data. *IEEE Access*, *9*, 109488–109496. https://doi.org/10.1109/ACCESS.2021.3102616

Pramita, D. A. K., Saraswati, N. W. S., Sandana, I. P. D., Pirozmand, P., & Bisena, I. K. A. (2024). Optimizing Hotel Room Occupancy Prediction Using an Enhanced Linear Regression Algorithms. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *24*(1), 95–104. https://doi.org/10.30812/matrik.v24i1.4254

Rahmaddeni, R., Wicaksono, M. T., Wulandari, D., Agustriono, A., & Ibrahim, S. A. (2024). Enhancing Multiple Linear Regression with Stacking Ensemble for Dissolved Oxygen Estimation. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *24*(1), 85–94. https://doi.org/10.30812/matrik.v24i1.4280

Ray, S., Rahman, M. M., Haque, M., Hasan, M. W., & Alam, M. M. (2023). Performance evaluation of SVM and GBM in predicting compressive and splitting tensile strength of concrete prepared with ceramic waste and nylon fiber. *Journal of King Saud University - Engineering Sciences*, *35*(2), 92–100. https://doi.org/10.1016/j.jksues.2021.02.009

Sampaio, C., Sebastião, J. R., & Farinha, L. (2024). Hospitality and Tourism Demand: Exploring Industry Shifts, Themes, and Trends. *Societies*, *14*(10), 1–14. https://doi.org/10.3390/soc14100207

Shirisha, N., Anusha, K., Kiran, A., & Buavani, Y. T. S. (2023). Prediction of Hotel Booking & Cancellation using Machine Learning Algorithms. *2023 International Conference on Computer Communication and Informatics (ICCCI)*, 1–4. https://doi.org/10.1109/ICCCI56745.2023.10128484

Soegianto, L. M., Hinandra, A. T., Suri, P. A., & Fajar, M. (2024). Comparison of Model Performance on Housing Business Using Linear Regression, Random Forest Regressor, SVR, and Neural Network. *Procedia Computer Science*, *245*(C), 1139–1145. https://doi.org/10.1016/j.procs.2024.10.343

Thomas, N. S., & Kaliraj, S. (2024). An Improved and Optimized Random Forest Based Approach to Predict the Software Faults. *SN Computer Science*, *5*(5). https://doi.org/10.1007/s42979-024-02764-x

Zhang, B., Li, N., Law, R., & Liu, H. (2022). A hybrid MIDAS approach for forecasting hotel demand using large panels of search data. *Tourism Economics*, *28*(7), 1823–1847. https://doi.org/10.1177/13548166211015515

Zhang, H., & Lu, J. (2022). Forecasting hotel room demand amid COVID-19. *Tourism Economics*, *28*(1), 200–221. https://doi.org/10.1177/13548166211035569

*name of corresponding author